Clinical Radiology 73 (2018) 827–831

Contents lists available at ScienceDirect

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net

Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification

E.J. Yates [a,*], L.C. Yates [a], H. Harvey [b]

[a] Foundation Doctor, West Midlands, England, UK
[b] Kheiron Medical Technologies, Rocketspace, 40 Islington High St, London N1 8EQ, UK

ARTICLE INFORMATION

Article history:
Received 27 February 2018
Accepted 9 May 2018

AIM: To develop a machine learning-based model for the binary classification of chest radiography abnormalities, to serve as a retrospective tool in guiding clinician reporting prioritisation.

MATERIALS AND METHODS: The open-source machine learning library, Tensorflow, was used to retrain a final layer of the deep convolutional neural network, Inception, to perform binary normality classification on two, anonymised, public image datasets. Re-training was performed on 47,644 images using commodity hardware, with validation testing on 5,505 previously unseen radiographs. Confusion matrix analysis was performed to derive diagnostic utility metrics.

RESULTS: A final model accuracy of 94.6% (95% confidence interval [CI]: 94.3–94.7%) based on an unseen testing subset (n=5,505) was obtained, yielding a sensitivity of 94.6% (95% CI: 94.4–94.7%) and a specificity of 93.4% (95% CI: 87.2–96.9%) with a positive predictive value (PPV) of 99.8% (95% CI: 99.7–99.9%) and area under the curve (AUC) of 0.98 (95% CI: 0.97–0.99).

CONCLUSION: This study demonstrates the application of a machine learning-based approach to classify chest radiographs as normal or abnormal. Its application to real-world datasets may be warranted in optimising clinician workload.

© 2018 The Royal College of Radiologists. Published by Elsevier Ltd. All rights reserved.

Introduction

Increasing healthcare demand is this reflected in radiology departments and subsequent clinician workload.[1] Despite radiological advances, the chest radiography (CXR) remains the most commonly requested imaging technique in the UK.[2] Consequently, timely radiologist reporting of every film is not always possible, leading to a "backlog" of unreported studies. In order to deliver maximal patient benefit from this reporting backlog, a system of image abnormality prioritisation would be beneficial, allowing reporting to first focus on examination of pathology over normality.

The "red dot" system, a method of radiographer communication of potential image abnormality has been in practice for almost 40 years.[3] The modern system uses digital superimposition of the words "red dot" on such images, a tribute to the traditional method of affixing a circular, red sticker to the abnormal plain film. Systematic review of radiographer red dot usage found 78% (74–82%) sensitivity and 91% (88–93%) specificity across pooled chest and abdominal films, highlighting the role of triage.[4]

* Guarantor and correspondent: E. Yates, Foundation Doctor, West Midlands, England, UK.
E-mail address: elliotyatesj@gmail.com (E.J. Yates).

https://doi.org/10.1016/j.crad.2018.05.015
0009-9260/© 2018 The Royal College of Radiologists. Published by Elsevier Ltd. All rights reserved.

Although radiographers can be trained in this prospective image prioritisation methodology, it does not tackle the retrospective burden of imaging backlog.

Machine learning (ML), specifically the field of image recognition using neural networks, could provide one such avenue of pictorial classification. Deep convolutional neural networks (CNN), an architecture loosely modelled on the biological organisation of the human brain (Fig 1), represents one such machine learning approach, historically demonstrating breakthroughs in computer vision and speech recognition.[5]

Such methods have previously been applied to radiology, with advances in multiple imaging techniques. CheXNet,[6] a CNN trained to yield both probability and heat map localisation of pneumonia in chest radiographs, demonstrated an F1 score of 0.435 (95% confidence interval [CI]: 0.387–0.481), which was a statistically significant improvement on a pooled radiologist average (0.387, 95% CI: 0.330–0.442). Furthermore, Cicero *et al.* retrained the GoogLeNet CNN in the multi-label classification of chest radiograph pathology, demonstrating a peak 91% sensitivity and specificity for a single disease category.[7]

As discussed, existing ML approaches have typically focused on formal reporting of pathology or multi-label classification, rather than prioritisation based on abnormality. The aim of the present study was to explore the ability of a deep-learning, computer vision approach in binary normality classification of plain film chest radiographs to serve as a rapid screening tool to assist clinicians in prioritising scans for formal reporting.

Materials and methods

The National Institutes of Health (NIH) Clinical Center 'ChestX-ray14' chest radiography corpus[8] includes 112,120 images derived from more than 30,000 patients. The anonymised dataset of frontal-view images with corresponding labels (text-mined into 14 disease categories from the corresponding original radiology report) represents a valuable resource for neural network training. The thoracic pathology categories include: atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, oedema, emphysema, fibrosis, pleural thickening, and hernia (Fig 2).

Additionally, the Indiana University hospital network chest radiograph database,[9] previously the largest public dataset before the NIH release, contains 7,470 frontal and lateral view images with corresponding full radiologist narrative report. Both datasets were combined to facilitate neural network training across both normal and abnormal images (Fig 3).

Image preparation and metadata parsing

Images were obtained from the two aforementioned public repositories. Indiana University images were manually separated into frontal (n=3,819) and lateral (n=3,651)

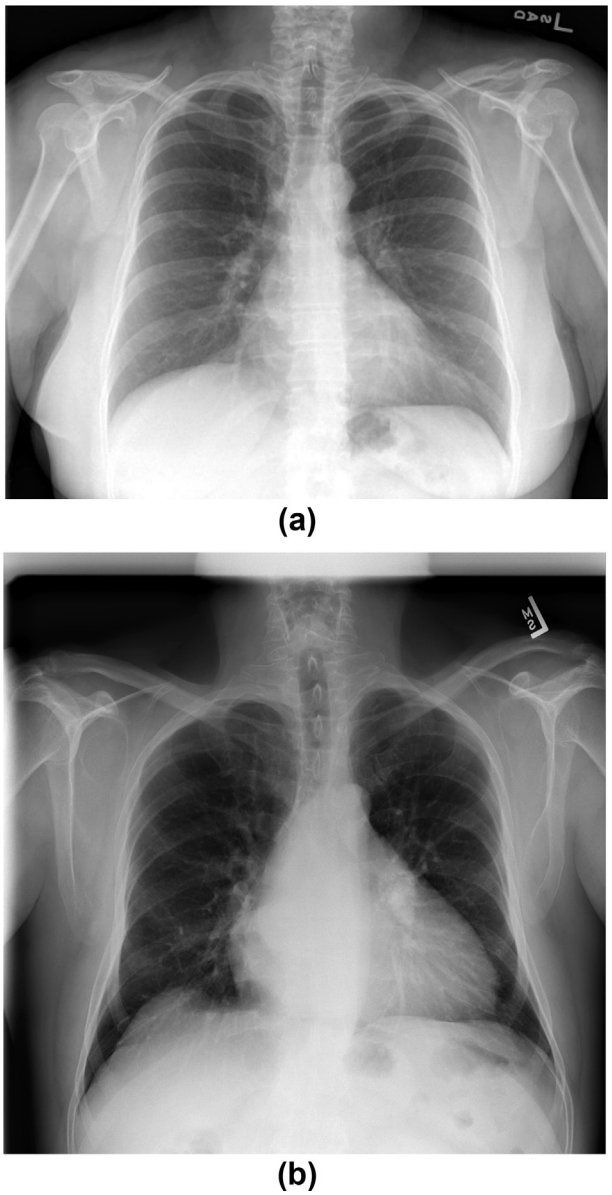

Figure 2 (a) Example of a "normal" image from the Indiana University Chest X-ray Collection. Image reference: *CXR100_IM-0002-1001*. CC BY-NC-ND 4.0 licence. (b) Example of an "abnormal" (cardiomegaly, emphysema) image from the ChestX-ray14 corpus. Image reference: *00000001_001*. Licence usage unrestricted.

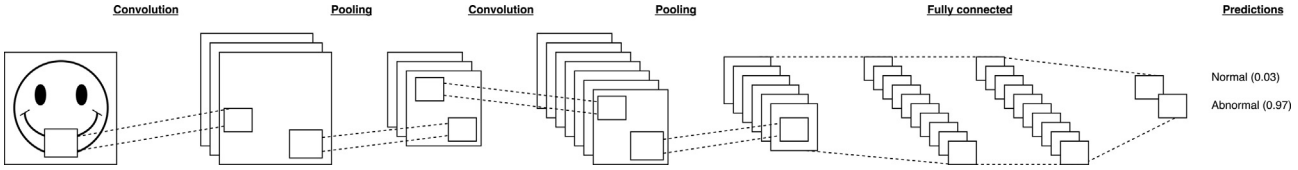

Figure 1 A pictorial representation of the "layers" of a deep CNN.

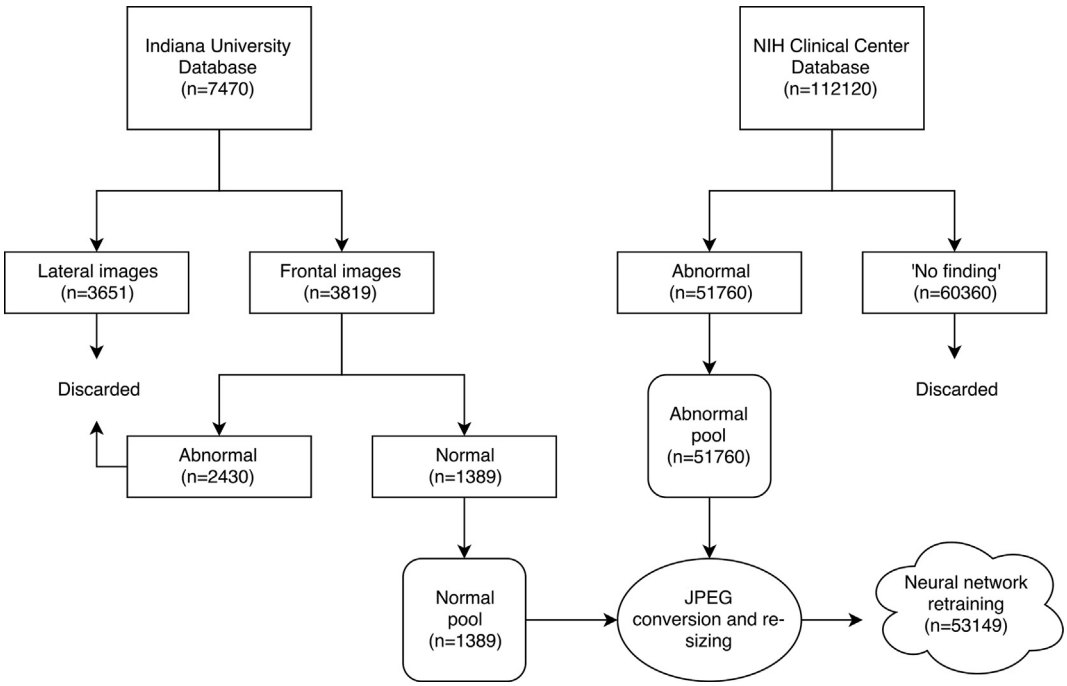

Figure 3 Flowchart of image pre-processing and binary classification.

groups (by two independent hospital doctors with 1 year of experience, neither with formal radiology training), with the lateral images being discarded.

NIH images were classified as "abnormal" if any of the aforementioned 14 common thoracic disease categories were present (n=51,760) in the corresponding metadata report. Images labelled as "no finding" (n=60,360) did not contain any of the 14 categories; however, these were not guaranteed to be normal and such images were consequently discarded. Indiana University images bearing the major subfield free-text "normal" were classified as "normal", all other images were discarded.

All included images were converted to greyscale JPEG format and re-sized (without aspect ratio preservation) to 299×299 pixels using Imagemagick (version 7.0.7–21). No additional image transformations or distortions were performed. The final training corpus (n=53,149) of frontal images was divided into "normal" (n=1,389) and "abnormal" (n=51,760) folders ahead of model training.

Model selection and neural network transfer learning

Tensorflow (version 1.4.1), an open-source machine learning library developed by Google[10] was used as a framework for retraining the existing deep convolutional neural network, Inception, version 3[11]; a model honed for image recognition. From scratch CNN training without graphics processing unit (GPU) support would typically take weeks; consequently, transfer learning was used to shortcut this computationally intensive task by retraining the Inception model to classify plain film radiographs on only the final "layers" of the neural network. The Tensorflow online documentation[12] provides an introduction to machine learning for beginners.

A modified Tensorflow image classification fine-tuning script,[13] capable of yielding softmax weights in addition to final confusion matrix parameters, was used to fine tune Inception_v3 (inception-2015-12-05). The following hyperparameters were used; learning rate of 0.01, batch size of 100, and 10,000 training steps. The original corpus was separated based on a training-to-testing ratio of 90:10%. This allocated n=47,644 images for initial transfer learning, with an unseen testing subgroup of n=5,505 images. Additionally, a cross-validation set of 10% (batch size of 100) was used during training. All data handling was performed on a single commodity cloud server (16 GB memory, 8 vCPUs) without GPU support.

Statistical analysis

Sensitivity, specificity, and associated metrics were calculated using the online resource Statpages.[14] Uncertainty estimates were to the 95% CI. Receiver operating characteristic (ROC) curves were plotted using easyROC.[15]

Results

Image preparation, metadata parsing, and neural network retraining was performed in under 6-hours on a single commodity cloud server. No image transformations or distortions (typically used in machine learning image classification to augment the training dataset) were used as this was deemed to be an unlikely reflection of real-world radiology image datasets.

A final model accuracy (defined by overall fraction correct; Fig 4) of 94.6% (95% CI: 94.3–94.7%) based on an unseen testing subset (n=5,505) was obtained. This yielded sensitivity 94.6% (95% CI: 94.4–94.7%) and specificity 93.4%

$$\frac{True\ positives\ +\ True\ negatives}{Total}$$

Figure 4 Formula used for accuracy (defined by overall fraction correct).

(95% CI: 87.2–96.9%) with a positive predictive value (PPV) of 99.8% (95% CI: 99.7–99.9%) and negative predictive value (NPV) 27.9% (95% CI: 26–28.9%), all based on a cut value of 0.5 (Table 1). The ROC curve (Fig 5) demonstrated an area under the curve (AUC) of 0.98 (95% CI: 0.97–0.99). Cut-offs were unadjusted from 0.5.

Discussion

Demonstration of a machine learning "red dot" model

A "red dot" system is based on the recognition of suspected abnormality. The results obtained from this study demonstrate the model's ability as a machine learning "red dot" system as evidenced by the high overall fraction correct and favourable ROC curve characteristics.

The remit of this study was to prioritise abnormality detection and thus the safety of normality determination was not fully considered; consequently, the absence of a "red dot" gives no promise of normality. The model's false-positive rate (FPR) was between 3–13%. Previous meta-analyses of radiographer "red dot" usage in combined chest and abdominal radiograph interpretation in the emergency department setting identified FPRs between 7–12%,[4] which suggests reasonable concordance with existing literature from real-world settings. Additionally of note, model abnormality detection gives neither weight to the severity of the recognised disease process, nor any pointers to the suspected pathology.

Table 1
Confusion matrix from unseen testing subset.

	Ground truth abnormal	Ground truth normal
Model predicted abnormal	5092	8
Model predicted normal	292	113

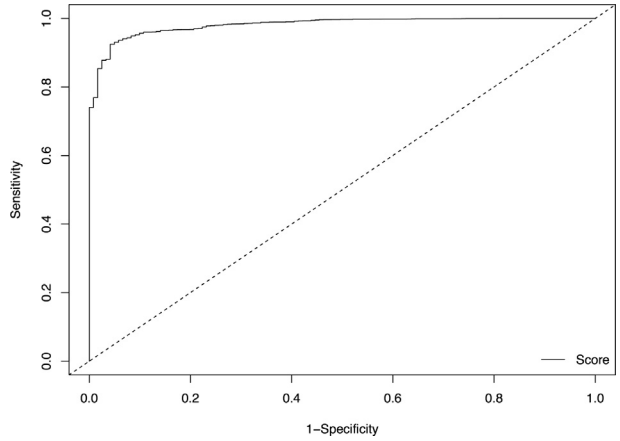

Figure 5 ROC curve for retrained model.

Projected effective clinician time utilisation

In the proposed system of clinician workload optimisation through preferential reporting of predicted abnormal studies, the benefits would only be recognised if the selected images demonstrated true abnormality. The almost 100% PPV represents a strength of this model in presenting such cases. Adjusted PPVs based on pre-test probability were not possible given a lack of knowledge of the underlying training corpus unselected epidemiology.

The possibility for expansion of this model into high-throughput clinical areas (e.g., emergency departments) across a variety of imaging techniques is an exciting avenue for future development. Therein, in addition to radiologist time optimisation, a multi-modal "red dot" may guide junior clinicians to previously unrecognised abnormality, when applied to imaging techniques not routinely formally reported.

Limitations of the source data

This work represents an open-source development on a publically available dataset. ChestX-ray14 is the largest public collection of chest radiographs, which are freely available online without significant licensing restrictions. This example of anonymous, large-scale data sharing is a highly valuable resource for data science, specifically machine learning. Expansion of such repositories from a variety of healthcare settings would provide further training material for machine learning models, thereby fuelling innovation in patient diagnostics and system efficiency.

The ChestX-ray14 corpus ground truth labels were text-mined, using natural language processing, from the corresponding full radiological report. Such labels therefore have no reference-standard human confirmation; however, the text-mining model was validated on the aforementioned Indiana University reports, achieving an F1-score of 0.90.[8] This lack of external corpus validation has been challenged,[16] proposing that compared to human inspection, the ChestX-ray14 text-mined ground truth labels are inaccurate. Consequently, in the absence of full radiology report release with the corpus, optimisation of ground truth accuracy will likely require significant radiologist re-reporting.

Furthermore, the dataset provides no details on the setting of data collection (inpatient or outpatient), the imaging indication (screening, acute admission, or trauma), the expected range of abnormalities outside of the 14 disease labels (e.g., pacemakers, invasive lines) or on normal clinical variations (e.g., rotation, kyphosis). Consequently, this lack of diversity guarantee may limit generalisability to a heterogeneous real-world case mix.

Limitations of the model

Overfitting (as seen in the classic tank recognition story of image classification)[17] is a problem whereby images of the same label share features that a human would know is

not related to the label. Although this dataset has been internally tested on >5,000 images, this does not represent a real-world validation. As such, further research is required to validate the application of such models to non-publicly available research datasets.

CNNs have typically been used in multi-label image classification; this model demonstrated binary classification. Consequently, the vulnerability to within-class appearance variation represents a challenge to the generalisability of this model to real-world datasets. Ideally, model "misclassifications" would be used to identify systematic model inaccuracies. Although it may be interesting to see examples of incorrectly labelled radiographs; without larger testing subsets, such cases would be randomly selected, and are, therefore, unlikely to represent true systematic model inaccuracies.

Finally, due to image availability, the model was trained on mostly abnormal chest radiographs (97%). This epidemiology of abnormality is highly unlikely to represent a real-world case mix, with abnormality in a previous similar study corpus accounting for approximately one-third of the total cases.[7] The combination of training subset abnormality imbalance and model optimisation for high PPV likely accounts for the greater false-negative over FPR; the model has more examples of abnormality and therefore a greater predictive ability.

In conclusion, this study sought to apply machine learning to the binary normality classification of plain film chest radiographs, in order to assist clinicians in the prioritisation of formal reporting. A "red dot" marker of abnormality, similar to the traditional radiographer system, was designed based on the retraining of deep CNNs to the process of computer vision.

The model, based on the Inception version 3 CNN was trained on almost 50,000 anonymised chest radiographs using commodity hardware, without GPU acceleration, in a matter of hours. The yielded model demonstrated high accuracy and PPV, highlighting the model's ability to detect abnormality. The high PPV ensures optimal clinician time utilisation by selecting abnormality based on 14 key thoracic pathologies.

Although further work is required to validate the application of such models to real-world datasets, the present study adds to existing literature in proposing the application of deep machine learning to the rapid automatic detection of abnormality in radiological imaging. Machine learning-based screening on radiology datasets may therefore have an exciting future in optimising clinician workload in the face of expanding demand.

References

1. Grant L, Appleby J, Griffin N, *et al*. Facing the future: the effects of the impending financial drought on NHS finances and how UK radiology services can contribute to expected efficiency savings. *Br J Radiol* 2012 Jun;**85**(1014):784–91.
2. Health Protection Agency. *Frequency of medical and dental X-ray examination in the UK 2008*. Oxfordshire: HPA; 2010.
3. Berman L, de Lacey G, Twomey E, *et al*. Reducing errors in the accident department: a simple method using radiographers. *Br Med J (Clin Res Ed)* 1985 Feb 9;**290**(6466):421–2.
4. Brealey S, Scally A, Hahn S, *et al*. Accuracy of radiographers red dot or triage of accident and emergency radiographs in clinical practice: a systematic review. Clin Radiol. 2006 Jul;61(7):604-615.
5. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. *RadioGraphics* 2017 Mar-Apr;**37**(2):505–15.
6. Rajpurkar P, Irvin J, Zhu K, *et al*. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225v3 [cs.CV].
7. Cicero M, Bilbily A, Colak E, *et al*. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Invest Radiol* 2017 May;**52**(5):281–7.
8. Wang X, Peng Y, Lu L, *et al*. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, 21–26 July*. New York, NY: IEEE; 2017.
9. Demner-Fushman D, Kohli MD, Rosenman MB, *et al*. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc* 2016 Mar;**23**(2):304–10.
10. Abadi M, Barham P, Chen J, *et al*. TensorFlow: a system for large-scale machine learning. arXiv:1605.08695.
11. Szegedy C, Vanhoucke V, Loffe S, *et al*. Rethinking the inception architecture for computer vision. arXiv:1512.00567.
12. TensorFlow. Getting started. Available at: https://www.tensorflow.org/get_started/. Accessed 17 April 2018.
13. Tensorflow. GitHub. Available at: https://github.com/tensorflow/tensorflow. Accessed 17 April 2018.
14. Statpages.info. Interactive statistical calculation pages. Available at: http://statpages.info/index.html. Accessed 15 January 2018.
15. Goksuluk D, Korkmaz S, Zararsiz G, *et al*. easyROC: an interactive web-tool for ROC curve analysis using R language environment. *R J* 2016;**8**(2):213–30.
16. Oakden-Rayner L. Exploring the ChestXray14 dataset: problems. Wordpress.com. Available at: https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/. Accessed 20 February 2018.
17. Kaufman J. Detecting tanks. Jefftk.com. Available at: https://www.jefftk.com/p/detecting-tanks. Accessed 15 January 2018.